ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 441-446

Development of a French Speech Recognizer Using a Hybrid HMM/MLP System

Jean-Marc Boite and Christophe Ris

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez
B-7000 Mons, Belgium
Email: boite@tcts.fpms.ac.be

Abstract. In this paper we describe the development of a French speech recognizer, and the experiments we carried out on our hybrid HMM/ANN system which combines Artificial Neural Networks (ANN) and Hidden Markov Models (HMMs). A phone recognition experiment with our baseline system achieved a phone accuracy of about 75% which is very similar to the best results reported in the literature [1]. Preliminary experiments on continuous speech recognition have set a baseline performance for our hybrid HMM/ANN system on BREF using lexicons of different sizes. All the experiments were carried out with the STRUT (Speech Training and Recognition Unified Toolkit) software [2] and the NOWAY large vocabulary decoder [3]

1. Introduction

Significant advances have been made in recent years in the area of large vocabulary speaker independent continuous speech recognition. Hidden Markov Models (HMMs) are nowadays the most successful modelling approach for speech recognition. A good introduction to HMMs and their use in speech recognition tasks can be found in [4]. In a classical HMM framework (see figure 1), probabilities are usually estimated by mixtures of gaussians: the estimated probability is a weighted sum of normal density functions. Vector quantization can also be used, and discrete probability density functions are estimated. Once the probabilities are obtained, a dynamic programming is performed to find the best path in the HMM, using the Viterbi algorithm [5].

However, standard HMMs suffer from strong assumptions among which the observation independence assumption stating that the acoustic vectors are not time correlated, or the underlying HMM state distribution assumption. These assumptions can be relaxed by introducing neural networks in the HMM framework [6] [1]. The neural network (a multi-layer perceptron in our particu-

[1] known as hybrid HMM/ANN method

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 441-446

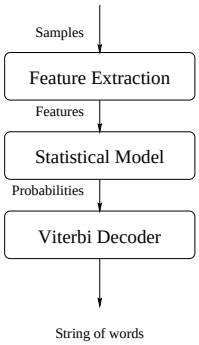

Figure 1: *Automatic Speech Recognizer*

lar case) estimates the state posterior probabilities which will be used by the HMMs. The hybrid HMM/ANN system has already been successfully applied in American English and British English. In this paper, this system is tested on a French continuous speech database. BREF-80 [7] is a large read speech corpus from 80 speakers. The text material was selected from the French newspaper *Le Monde* so as to provide a large vocabulary (over 20,000 words) and a wide range of phonetic environments. As BREF contains 1115 distinct diphones and over 17,500 triphones, it can be efficiently used to train phonetic models. Some good results on this particular database have already been published [8, 1] using continuous density HMMs and context dependent models. We will compare those results with our baseline context independent hybrid system on this particular task.

2. Acoustic Features

Prior to the application at the input of a neural network, the speech utterance must be analysed. A window, usually of 20 or 30 ms is applied to the input samples. A frequency analysis is performed, and acoustic features are extracted. Four sets of acoustic features have been used: the Perceptual Linear Predictive coefficients (PLP), the log-RASTA-PLP coefficients [9], the lpc-cepstral features with cepstral mean subtraction (CMS) [10] and the mel-scale frequency cepstral coefficients (MFCC). The feature set for our hybrid HMM/ANN systems was based on a 26 dimensional vector composed of the raw *acoustic* parameters, the Δ*parameters*, the Δ*energy* and the $\Delta\Delta$*energy*. The input layer of the MLP received nine frames of contextual information were used at the input of the ANN, leading to 234 inputs. Finally, the hidden layer of our MLPs counts 1000 nodes, leading to approximately 270,000 parameters which is much lower than most of the LVCSR systems described in the literature. Larger MLPs (up to 8000 hidden nodes) have been trained sucessfully [16], thanks to highly efficient hardware developed for that purpose [17]. But since our database was limited, we decided to stick to 1000 nodes.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 441-446

3. Labeling and Training Procedure

To train a hybrid HMM/ANN speech recognition system, we embed the training of the ANN in an iterative process. Using the Viterbi algorithm [5] and the statistical model we have obtained so far, we generate a segmentation of the input speech signal. In other words, we assign a phonetic label to each segment of speech present in the database. These phonetic labels define targets for the ANN, which is trained with several passes (usually 7 or 8) through the database. The trained ANN is then used to generate a new segmentation. Three or four iterations are generally performed.

The problem with this method is to get the first segmentation. One of the first database that have been made available to the speech research community, TIMIT [11], has been hand labeled. Experts have been looking at the spectrogram of each utterance, and manually labeled the speech samples. That database is still widely used to boostrap the training of speech recognition systems in American English. Another method is to linearly segment the speech: every phonemes are considered to have the same duration, or, better, a duration specific to each phoneme: vowels are usually longer than consonants, for instance. While this method is valid for isolated words, where the utterance is short, it is not very practical for continuous speech, where complete sentences have to be segmented. In the development of our French recognizer, we used a high-quality digital speech synthesizer (MBROLA [12]) to create, from the phonetic transcription, a reference speech pattern with known phoneme boundaries and then align the natural speech on this pattern [13]. So the alignment process is reduced to a simple dynamic time warping (DTW).

While this approach apparently looses speaker independence, it turned out that the much better segmental information counterbalances the dependence on a single reference voice. Nevertheless, two different voices (a male and a female voice) significantly improve the relevance of the segmentation. Malfrère and Dutoit reported in [13] a segmentation error rate of about 8% (assuming correct a time deviation lower than 50 ms against a manual segmentation).

After four iterations of forced Viterbi alignments, we obtained around 80% recognition rate at the frame level. In other words, given 9 frames of acoustic features (90 ms of speech), the MLP was able to recognize the right phoneme eight times out of ten.

4. Experiments

A phone recognition experiment has been carried out on the BREF corpus. The baseline phone recognizer uses a set of 35 CI (Context Independent) phone models. Each phoneme model is a N identically distributed states, left to right HMM where N is estimated from the mean duration of each phoneme, and No grammar constraints were used. A phone accuracy of 75% has been obtained, which is very similar to the best results reported in the literature [1], but using a much simpler system.

ESANN'1999 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 21-23 April 1999, D-Facto public., ISBN 2-600049-9-X, pp. 441-446

The phone recognition rates obtained on the BREF database motivated our experiments at the word level. The same test set as for the phone recognition experiments has been used with different vocabulary sizes: 1K, 3K, 13K and 64,000 words. The phonetic transcriptions were produced by a rule based phonetizer. In addition, we used a bigram (perplexity 151.6) and a trigram (perplexity 94.4) language models estimated on texts extracted from the French newspaper "Le Monde" (1990-1992, ∼80M words) using the CMU-Cambridge SLM toolkit [14].

The results obtained with our baseline CI hybrid HMM/ANN system (using the Neural Network trained for the Phoneme Recognition Experiment) are reported in table 1. All those experiments were performed with PLP features.

Dictionary Size	1K	3K	13K	64K
Error rate	15.9%	19.8%	24.1%	25.0%

Table 1: *Word Error rates using a classical hybrid HMM/ANN system and PLP features on different vocabulary sizes - Trigram language model*

Note that we encountered some difficulties inherent in the French language such as the liaisons between the words, the elision of some phonemes in particular phonetic contexts. Those problems were partially solved using multiple pronunciations. But the main problem in French are the homophones, as well single word than multiple words homophones. Gauvain [1] reports a homophone rate of 30% for French against only 3% in English (TIMIT).

The recognition rates at the phoneme level as well as a detailed study of the errors made by the system on continuous speech tasks show that the acoustic modelization of the speech signal is quite accurate. To improve our word recognition rate, we focussed on the pronunciation dictionaries and the language model. We first took care of the pre-processing of the text corpus, taking into account most frequent compound words, filtering syntactically incorrect sentences, etc. We extended the text corpus to years 96-98, leading to over 140 million words. We manually checked most of the phonetic transcriptions, and we applied an automatic pronunciation learning tool to the database. With all these improvements, we were able to lower the error rate to 18.5% on the 64 K words task. On another side, as Gauvain and Al. [1] reported an error reduction of 14% by using context dependent (CD) models compared to their best CI models, we also intend to test context dependent phone models in a hybrid framework [15]. However the rather limited amount of available data could be a limitation for the development of such a system.

5. Conclusion

These preliminary experiments have set a baseline performance for our hybrid HMM/ANN system on BREF. The phone recognition performance is similar to

the best results reported in the literature, using a much simpler system: context independent phonemes, no phone syntax, less parameters than the continuous densities HMM approaches.

At the word level, the best result we got was 18.5% word error rate on BREF. To our knowledge, this is the second best system on that database. The best one is around 12%, and is using multi-gaussian modelling. However, the development of that system took around ten years, and demands much more CPU.

Throughout all these experiments, we have found that MLPs are fairly easy to train, once you have the tools to do it. So, they not only lead to good results with the current state of the art, but they are very flexible, and open to new ideas and new theories. They allowed us to explore new fields, such as multi-band speech recognition [18] or mixture of experts [19].

References

[1] Gauvain J.L., and Lamel L., *"Experiments on speaker-independent phone recognition using BREF"*, Proceedings ICASSP'92, pages 557-560, San Francisco, March 1992.

[2] *"Step by Step guide to using the Speech training and recognition unified Tool (*STRUT*)"*, Available via ftp at *http://tcts.fpms.ac.be/speech/strut/users-guide/users-guide.html.*

[3] Cook G. D., Kershaw D. J., Christie J. D. M., Seymour C. W., and Waterhouse S. R. *"Transcription of broadcast television and radio news: The 1996 Abbot system"* Proceedings ICASSP'97, pp. 723-726, Munich, Germany, 1997.

[4] A.B. Poritz, *"Hidden Markov Models: a guided tour"*, Proceedings Int.Conf. Acoustics, Speech and Signal Processing, pages 7–13, New-York, 1988.

[5] G.D. Forney, Jr., *"The Viterbi Algorithm"*, Proceeding of the IEEE, vol 61, pp. 268–278, 1973.

[6] Bourlard H. and Morgan N., *"Connectionist Speech Recognition - A Hybrid Approach"*, Kluwer Academic Publisher, 1994.

[7] Lamel L.F., Gauvain J.L. and Eskénazi M., *"BREF, a Large Vocabulary Spoken Corpus for French"*, Proceedings EuroSpeech 1991, pp. 505-508, Geneva, Italy

[8] Young S.J. and al, *"Multilingual large vocabulary speech recognition : the European* SQALE *project"*, Computer Speech and Language 0-73-89-0, 1997.

[9] Hermansky H. and Morgan N., *"RASTA processing of speech"*, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4 pp. 578-589, 1994.

[10] Van Hamme H., Gallopyn G., Van Springel W., D'Hoore B., Butnaru M. and Bourlard H., *"Acoustic Features Comparison and Robustness Tests of a Realtime Recognizer on a Hardware Telephone Line Simulator"*, Proceedings of ICSLP 1994.

[11] John S. Garofolo, *"Getting Started with the DARPA TIMIT CD-ROm : An acoustic Phonetic Continuous Speech Database"*, National Institute of Standards and Technology (NIST), 1988

[12] Dutoit T., Pagel V., Pierret N., Bataille F. and Van Der Vreken O., *"The MBROLA Project : Toward a Set of High quality Speech Synthesizers Free of Use for non commercial purposes"*, Proceedings ICSLP'96, pp. 1393-1396, Philadelphia, PA, USA, 1996.

[13] Deroo O., Malfrere F. and Dutoit T., *"Phonetic Alignment : Speech Synthesis Based vs. Hybrid HMM/ANN"*, Proceedings EUSIPCO 1998, Rhodes, Greece.

[14] P.R. Clarkson and R. Rosenfeld. *"Statistical Language Modeling Using the CMU-Cambridge Toolkit"*. Proceedings EUROSPEECH'97, Rhodes, Greece, 1997.

[15] Dupont S., Ris C., Deroo O., Fontaine V., Boite J.M., and Zanoni L., *"Context Independent and Context Dependent Hybrid HMM/ANN Systems For Vocabulary Independent Tasks"*, Proceedings EUROSPEECH'97, pp. 1947-1950, Rhodes, Greece, 1997.

[16] Dan Ellis and Nelson Morgan, *"Size matters: An Empirical Study of Neural Network Training for Large Vocabulary Continuous Speech Recognition"*, To appear in Int.Conf. Acoustics, Speech and Signal Processing, 1999.

[17] John Wawrzynek, Krste Asanović, Brian E. D. Kingsbury James Beck, David Johnson, and Nelson Morgan, *"SPERT-II a vector microprocessor system"*, IEEE Computer, 29(3):79–86, March 1996.

[18] H. Bourlard and S. Dupont, *"A new ASR approach based on independent processing and recombination of partial frequency bands"*, Proc. of Intl. Conf. on Spoken Language Processing, (Philadelphia), pp 422–425, Oct 1996.

[19] Steve Waterhouse and Tony Robinson, *"Non-Linear Prediction of Acoustic Vectors Using Hierarchical Mixtures of Experts"*, Neural Information Processing Systems 7, Morgan Kaufmann 1994.